Testing interval forecast: a GMM approach

Elena-Ivona Dumitrescu

Corresponding author. Maastricht University and LEO, Université d'Orléans. Faculté de Droit, d'Economie et de Gestion. Rue de Blois - BP 6739. 45067 Orléans Cedex 2.
Email: elena.dumitrescu@univ-orleans.fr.

Christophe Hurlin

LEO, Université d'Orléans. Email: christophe.hurlin@univ-orleans.fr.

Jaouad Madkour

LEO, Université d'Orléans. Email: jaouad.madkour@univ-orleans.fr.

Abstract

This paper proposes a new evaluation framework of interval forecasts. Our model free test can be used to evaluate intervals forecasts and/or and High Density Region, potentially discontinuous and/or asymmetric. Using simple J-statistic based on the moments defined by the orthonormal polynomials associated with the Binomial distribution, this new approach presents many advantages. First, its implementation is extremely easy. Second, it allows for a separate test for unconditional coverage, independence and conditional coverage hypothesis. Third, Monte-Carlo simulations show that for realistic sample sizes, our GMM test outperforms traditional LR test. These results are corroborated by an empirical application on SP500 and Nikkei stock market indexes. An empirical application for financial returns confirms that using GMM test leads to major consequences for the ex-post evaluation of interval forecasts produced by linear versus non linear models.

Key words: Interval forecasts, High Density Region, GMM. JEL codes : C52, G28.

1 Introduction

In recent years, the contribution of nonlinear models to forecasting macroeconomic and financial series has been intensively debated (see Teräsvirta, 2006, Colletaz and Hurlin, 2005 for a survey). A suggested by Teräsvirta, there exist relatively large studies in which the forecasting performance of nonlinear models is compared with that of linear models using actual series. In general, no dominant nonlinear (or linear) model has emerged. However, the use of nonlinear models has actually led to the renewal of the short-term forecasting approach, especially through the emergence of concepts like High Density Regions (Hyndman, 1995, thereafter HDR) or density forecasts as opposed to point forecasts. Consequently, this debate on non-linearity and forecasting involves the new forecast validation criteria. It is the case of density forecasts, for which many specific evaluation tests have been developed (Bao, Lee and Saltoglu, 2004, Corradi and Swanson 2006 etc.).

On contrary, if there are numerous methods to calculate HDR and interval forecasts (Chatfield, 1993), only a few studies propose validation methods adapted to these kind of forecasts. This paradox is even more astounding if we take into consideration the fact that interval forecast is the most generally used method by applied economists to account for forecast uncertainty.

One of the main exceptions, is the seminal paper of Christoffersen (1998), that introduces general definitions of hypotheses allowing to assess the validity of an interval forecast obtained by using any type of model (linear or nonlinear). His model-free approach is based on the concept of violation: a violation is said to occur if the *ex-post* realization of the variable does not lie in the *ex-ante* forecast interval. Three validity hypothesis are then distinguished. The *unconditional coverage* hypothesis means that the expected frequency of violations is precisely equal to the coverage rate of the interval forecast. The *independence hypothesis* means that if the interval forecast is valid then violations must be distributed independently. In other words, there must not have any cluster in the violations sequence. Finally, under the *conditional coverage hypothesis* the violation process satisfies the assumptions of a martingale difference. Based on these definitions, Christoffersen proposes a Likelihood Ratio (hereafter LR) test to each of these hypotheses, by considering a binary first-order Markov chain representation under the alternative hypothesis. However, this approach imposes that a invalid interval forecast generate a first order time-dependency amongst violations.

More recently, Clements and Taylor (2002) applied a simple logistic regression with periodic dummies and modified the first-order Markov chain approach in order to detect dependence at a periodic lag. In 2003, Wallis recast Christoffersen (1998)'s tests in the framework of contingency tables increasing users' accessibility to these interval forecast evaluation methods. Owing to his innovative approach, it became possible to calculate exact p-values for the LR statistics in small sample cases.

Beyond their specificities, the main common characteristic of these tests is that assessing the validity of interval forecasts comes down to testing a distributional assumption for the violation process. If we define a binary indicator variable that takes a value one in case of violation, and zero otherwise, it is obvious that under the null, of CC the sum of the indicators associated a sequences of interval forecasts follows a Binomial distribution.

On these grounds, we propose in this paper a new GMM approach to test the interval forecasts validity. Relying on the GMM framework of Bontemps and Meddahi (2005), we define simple J-statistics based on particular moments defined by the orthonormal polynomials associated with the Binomial distribution. A similar approach has been used by Candelon and al. (2010) in the context of the Value-at-Risk [1] backtesting. The authors test the VaR forecasts validity by testing the geometric distribution assumption for the durations observed between two consecutive VaR violations. Here, we propose a more general approach for all kind of intervals and HDR forecasts, that directly exploits the properties of the violation process (and not the durations between violations). We adapt the GMM framework to the case of discrete distributions and more exactly to a binomial distribution by using an approach similar to subsampling, which is actually based on blocks of indicator variables.

This forthright method is thus general and easy to implement since it is based on slicing the indicator variable into blocks. Our approach has several advantages. First, we develop an unified framework in which the three hypotheses of unconditional coverage, independence and conditional coverage are tested independently. Second, contrary to LR tests, this approach imposes no restrictions under the alternative hypothesis. Third, this GMM-based test is easy to implement and does not generate computational problems regardless of the sample size. Fourth, this approach is proved to be robust to the uncertainty of distributional parameters. Finally, some Monte-Carlo simulations indicate that for realistic sample sizes, our GMM test have good power properties when compared to LR tests.

The paper is structured as follows. Section 2 presents the general framework of interval forecast evaluation, while section 3 introduces our new GMM-based evaluation tests. In section 4 we scrutinize the finite-sample properties of the tests through Monte-Carlo simulations and in section 5 we propose an empirical application. Section 6 concludes.

[1] Recall that the Value-at-Risk can be interpreted as a one-sided and open interval.

2 General Framework

Formally, let $x_t, t \in \{1, ..., T\}$ be a sample path of time series x_t. Let denote $\left\{CI_{t|t-1}(\alpha)\right\}_{t=1}^{T}$ the sequence of *out-of-sample* interval forecasts for the coverage probability α, so that

$$\Pr[x_t \in CI_{t/t-1}(\alpha)] = \alpha. \tag{1}$$

Hyndman (1995) identifies three methods to construct a $100(1-\alpha)\%$ forecast region: (i) a symmetrical interval around the point forecast, (ii) an interval defined by the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the forecast distribution, (i) and a High Density Region (HDR). These three forecast regions are identical (symmetric and continuous) in the case of symmetric and unimodal distribution. By contrast, HDR_α is the smallest forecast region for asymmetric or multimodal distributions. When the interval forecast is continuous, $CI_{t|t-1}(\alpha)$ can be defined as in Christoffersen (1998), by $CI_{t|t-1}(\alpha) = [L_{t|t-1}(\alpha), U_{t|t-1}(\alpha)]$, where $L_{t|t-1}(\alpha)$ and $U_{t|t-1}(\alpha)$ are the limits of the *ex-ante* confidence interval for the coverage rate α.

Whatever the form of the HDR or the interval forecasts (symmetric or asymmetric, continuous or discontinuous), we define an indicator variable $I_t(\alpha)$, also called violation, as a binary variable that takes a value one if the realization of x_t does not belong to this region.

$$I_t(\alpha) = \begin{cases} 1, & x_t \notin CI_{t|t-1}(\alpha) \\ 0, & x_t \in CI_{t|t-1}(\alpha) \end{cases}. \tag{2}$$

Based on the definition of the violations process, a general testing criterion for interval forecasts can be established. Indeed, as stressed by Christoffersen (1998), the interval forecasts are valid if and only if the CC hypothesis is fulfilled, *i.e.* both the IND and UC hypotheses are satisfied.

Accordingly, under the IND hypothesis, violations observed at different moments in time (I_t, I_{t+i}, $\forall\ i \neq 0$) for the same coverage rate ($\alpha\%$) must be independent, *i.e.* we must not observe any clusters of violations. In other words, past violations should not be informative about the present or future violations. Under the UC assumption, the probability to have a violation is equal to the α coverage rate:

$$H_{0,UC} : \Pr[I_t(\alpha) = 1] = \mathbb{E}[I_t(\alpha)] = \alpha. \tag{3}$$

In other words, the UC property places a restriction on how often violations may occur, whereas the IND assumption restricts the order in which these violations may appear.

Christoffersen (1998) pointed out that in the presence of higher-order dynamics it is important to go beyond the UC assumption and test the conditional efficiency hypothesis (CC). Under the CC assumption, the conditional (on a past information set Ω_{t-1}) probability to observe a violation must be equal to the α coverage rate, *i.e.* the I_t process satisfies the properties of a difference martingale:

$$H_{0,CC} : \mathbb{E}[I_t(\alpha) \mid \Omega_{t-1}] = \alpha. \tag{4}$$

Consequently, a sequence of interval/HDR forecasts $\left\{CI_{t|t-1}(\alpha)\right\}_{t=1}^{T}$ has correct conditional coverage, if $\{I_t(\alpha)\}_{t=1}^{T}$ are *i.i.d.* and follows a Bernoulli distribution with a success probability equal to α:

$$I_t \underset{H_{0,CC}}{\sim} i.i.d.\ Bernoulli(\alpha),\ \forall t. \tag{5}$$

This feature of the violation process is actually at the core of most of the interval forecast evaluation tests (Christoffersen, 1998, Clements and Taylor, 2002, etc.) and so it is for our GMM-based test.

3 A GMM-Based Test

In this paper we propose an unified GMM framework for evaluating interval forecasts by testing the Bernoulli distributional assumption of the violation series $I_t(\alpha)$. Our analysis is based on the recent GMM distributional testing framework developed by Bontemps and Meddahi (2005) and Bontemps (2006). We first present the environment of the test, then we define the moment conditions used to test the interval forecasts efficiency, and finally we propose simple J-statistics corresponding to the three hypothesis of UC, IND and CC.

3.1 Environment Testing

Given the result (5), it is obvious that if the interval forecast has a correct conditional coverage, the sum of violations follows a Binomial distribution

$$H_{0,CC} : \sum_{t=1}^{T} I_t(\alpha) \sim B(T, \alpha) \tag{6}$$

A natural way to test CC, consists in testing this distributional assumption. However this property cannot be directly used, since, for a given sequence $\left\{CI_{t|t-1}(\alpha)\right\}_{t=1}^{T}$, we have only one observation for the sum of violations.

Figure 1. Partial sums y_h and block size N .

Hence, we propose here an original approach, derived from the subsampling methodology (Politis, Romano and Wolf, 1999), that transforms the violation series I_t into a new one, drawn from a Binomial distribution. More precisely, since under the null hypothesis the violations $\{I_t(\alpha)\}_{t=1}^{T}$ are independent, it is possible to cut the initial series of violations into H subsamples of size N, where $H = [T/N]$ (see Figure 1). Then, for each sub-sample, we define y_h, $h \in \{1, ..., H\}$ as the sum of the corresponding N violations:

$$y_h = \sum_{t=(h-1)N+1}^{hN} I_t(\alpha). \tag{7}$$

As a result, under the null hypothesis, the constructed processes y_h are *i.i.d.* $B(N, \alpha)$, and thus the null of CC that the interval forecasts are well specified can simply be expressed as follows:

$$H_{0,CC} : y_h \sim B(N, \alpha), \ \forall h \in \{1, ..., H\}. \tag{8}$$

3.2 Orthonormal Polynomials and Moment Conditions

There are many ways to test conditional coverage hypothesis through the distributional assumption (8). Following Bontemps and Meddahi (2005) and Bontemps (2006), we propose here to use a GMM framework based. The general idea is that for many continuous and discrete distributions, it is possible to associate some particular orthonormal polynomials whose expectation is equal to zero. These orthonormal polynomials can be used as moment conditions in a GMM framework to test for a specific distributional assumption. For instance, the Hermite polynomials associated to the normal distribution can be employed to built a test for normality (Bontemps and Meddahi, 2005). Another particular polynomials are used by Candelon and alii. to test for a geometric distribution hypothesis.

In the particular case of a Binomial distribution, the corresponding orthonormal polynomials are called Krawtchouk polynomials. These polynomials are defined as follows:

Definition 1 *The orthonormal Krawtchouk polynomials associated to a Binomial distribution $B(N,\alpha)$ are defined by the following recursive relationship*:

$$P_{j+1}^{(N,\alpha)}(y_h)=\frac{\alpha(N-j)+(1-\alpha)j-y_h}{\sqrt{\alpha(1-\alpha)(N-j)(j+1)}}P_j^{(N,\alpha)}(y_h)-\sqrt{\frac{j(N-j+1)}{(j+1)(N-j)}}P_{j-1}^{(N,\alpha)}(y_h), \tag{9}$$

where $P_{-1}^{(N,\alpha)}(y_h)=0$, $P_0^{(N,\alpha)}(y_h)=1$, and $j<N$.

Our test exploits these moment conditions. More precisely, let us define $\{y_1;...;y_H\}$ a sequence of partial sums defined by (7) and computed from the sequence of violations $\{I_t(\alpha)\}_{t=1}^{T}$. Under the null conditional coverage assumption, variables y_h are *i.i.d.* and have a Binomial distribution $B(N,\alpha)$, where N the subsample size. Hence, the null of CC can be expressed as follows:

$$H_{0,CC}:\ \mathbb{E}\left[P_j^{(N,\alpha)}(y_h)\right]=0,\quad j=\{1,..,m\}, \tag{10}$$

with $m<N$.

An appealing property of the test is that it allows to test separately for the UC and IND hypothesis. Let us remind that under the UC assumption, the unconditional probability to have a violation is equal to the coverage rate α. Consequently, under UC, the expectation of the partial sum y_h is then equal to αN, since:

$$\mathbb{E}(y_h)=\sum_{t=(h-1)N+1}^{hN}\mathbb{E}[I_t(\alpha)]=\alpha N,\ \forall h\in\{1,...,H\}. \tag{11}$$

Given the properties of the Krawtchouk polynomials, the null UC hypothesis can be expressed as

$$H_{0,UC}:\ \mathbb{E}\left[P_1^{(N,\alpha)}(y_h)\right]=0. \tag{12}$$

In this case, we need only to use the first moment condition defined by $P_1^{(N,\alpha)}(y_h)=(\alpha N-y_h)/\sqrt{N\alpha(1-\alpha)}$, since the condition $\mathbb{E}\left[P_1^{(N,\alpha)}(y_h)\right]=0$ is equivalent to the UC condition $\mathbb{E}(y_h)=\alpha N$.

Under the IND hypothesis, the $y_h(\alpha)$ series follows a $B(N;\beta)$ distribution, where the probability β is not necessarily equal to the coverage rate α. Thus, the IND hypothesis can simply be expressed as:

$$H_{0,IND}:\ \mathbb{E}\left[P_j^{(N,\beta)}(y_h)\right]=0\quad j=\{1,..,m\}, \tag{13}$$

with $m<N$. To sum up, on the one hand, when testing the CC hypothesis it is necessary to consider at least two moment conditions based on Krawtchouk polynomials and to fix the success probability to the coverage rate. On the other hand, if we test the UC hypothesis, only the first orthonormal polynomial

($P_1^{(N,\alpha)}$) is necessary, while when testing the IND hypothesis, the probability is not constrained anymore.

3.3 Testing Procedure

Let us define $P^{(N,\beta)}$ denotes a $(m,1)$ vector whose components are the orthonormal polynomials $P_j^{(N,\beta)}(y_h)$, for $j = 1, .., m$. Under the CC assumptions and some regularity conditions (Hansen, 1982) it can be shown that:

$$\left(\frac{1}{\sqrt{H}} \sum_{h=1}^{H} P^{(N,\alpha)}(y_h) \right)^{\perp} \Sigma^{-1} \left(\frac{1}{\sqrt{H}} \sum_{h=1}^{H} P^{(N,\alpha)}(y_h) \right) \underset{H \to \infty}{\overset{d}{\to}} \chi^2(m), \tag{14}$$

where Σ is the long-run variance-covariance matrix of $P^{(N,\alpha)}(y_h)$. By definition of orthonormal polynomials, this long-run variance-covariance matrix corresponds to the identity matrix. Therefore, the optimal weight matrix has not be estimated and the corresponding J-statistic is very easy to implement. Let us denote by $J_{CC}(m)$ the CC test-statistic associated to the $(m,1)$ vector of orthonormal polynomials $P^{(N,\alpha)}(y_h)$.

Definition 2 *Under the null hypothesis of conditional coverage, the CC test statistic verifies*:

$$J_{CC}(m) = \left(\frac{1}{\sqrt{H}} \sum_{h=1}^{H} P^{(N,\alpha)}(y_h) \right)^{\top} \left(\frac{1}{\sqrt{H}} \sum_{h=1}^{H} P^{(N,\alpha)}(y_h) \right) \underset{H \to \infty}{\overset{d}{\to}} \chi^2(m), \tag{15}$$

where $P^{(N,\alpha)}(y_h)$ is the $(m,1)$ vector of orthonormal polynomials $Pk_j^{(N,\alpha)}(y_h)$, for $j < m$.

Since the $J_{UC}(m)$ statistic corresponding to the UC hypothesis is a special case of the $J_{CC}(m)$ test statistic, it can be immediately computed by taking into account only the first moment condition $P_1^{(N,\alpha)}(y_h)$, and can be expressed as follows:

$$J_{UC} = J_{CC}(1) = \left(\frac{1}{\sqrt{H}} \sum_{h=1}^{H} P_1^{(N,\alpha)}(y_h) \right)^2 \underset{H \to \infty}{\overset{d}{\to}} \chi^2(1). \tag{16}$$

Finally, the independence hypothesis statistic, denoted $J_{IND}(m)$ takes the form of:

$$J_{IND}(m) = \left(\frac{1}{\sqrt{H}} \sum_{h=1}^{H} P^{(N,\beta)}(y_h) \right)^{\top} \left(\frac{1}{\sqrt{H}} \sum_{h=1}^{H} P^{(N,\beta)}(y_h) \right) \underset{H \to \infty}{\overset{d}{\to}} \chi^2(m), \tag{17}$$

where $P^{(N,\beta)}(y_h)$ is the $(m,1)$ vector of orthonormal polynomials $Pk_j^{(N,\beta)}(y_h)$ defined for a coverage rate β that can be different from α. The coverage rate β is generally unknown, and thus it has to be estimated. When using the squared-T-root-consistent estimator $\hat{\beta} = (1/T)\sum_t^T I_t(\alpha)$ instead of β, the asymptotic distribution of the $J_{IND}(m)$ test-statistic is likely to be modified, thus causing a distributional parameter uncertainty problem. Nevertheless, Bontemps (2006) proved that the asymptotic distribution does not change if the moments can be expressed as a projection onto the orthogonal of the score. More exactly, the maximum likelihood estimation of β implies that the score of maximum likelihood is equal to zero. Then, since the first moment $P_1^{(N,\hat{\beta})}$ is proportional to the score, this first moment is null by definition and only the degree of freedom of the GMM-statistic $J_{IND}(m)$ has to be adjusted accordingly:

$$J_{IND}(m) = \left(\frac{1}{\sqrt{H}}\sum_{h=1}^{H} P^{(N,\widehat{\beta})}(y_h)\right)^{\top}\left(\frac{1}{\sqrt{H}}\sum_{h=1}^{H} P^{(N,\widehat{\beta})}(y_h)\right) \underset{H\to\infty}{\overset{d}{\longrightarrow}} \chi^2(m-1), \tag{18}$$

where $P^{(N,\hat{\beta})}(y_h)$, is a $(m,1)$ vector, whose components are the orthonormal polynomials $Pk_j^{(N,\hat{\beta})}(y_h)$ defined for the estimated coverage rate $\widehat{\beta}$. The demonstration of this property in the particular case of Krawtchouk polynomials can be found in appendix A. At the same time, it can be noted that the other two tests are free of parameter uncertainty since the coverage rate α is defined *a-priori*.

4 Monte-Carlo Experiments

In this section we gauge the finite sample properties of our GMM-based test using Monte-Carlo experiments. We first analyze the size performance of the test and we then investigate its empirical power under various alternatives. A comparison with Christoffersen (1998)'s LR tests is provided for both analyses. In order to control for size distortions, we use Dufour (2006)'s Monte-Carlo method.

4.1 Empirical Size Analysis

To illustrate the size performance of our UC and CC tests in finite sample, we generate a sequence of T violations by taking independent draws from a Bernoulli distribution, considering successively a coverage rate $\alpha = 1\%$ and $\alpha = 5\%$. Several sample sizes T ranging from 250 (which roughly corresponds to one year of daily forecasts) to $1,500$ are considered. The size of the blocks (used to compute the H partial sums y_h) is fixed to $N = 25$ or $N = 100$ observations. Additionally, we consider several moment conditions m from 1 (for the UC test statistic J_{UC}) to 5. Based on a sequence $\{y_h\}_{h=1}^{H}$, with $H = [T/N]$, we compute both statistics J_{UC} and $J_{CC}(m)$. The reported empirical sizes correspond to the rejection rates calculated over 10,000 simulations for a nominal size equal to 5%.

[Insert Table 1]

In table 1, the rejection frequencies for the $J_{CC}(m)$ statistic and a block size N equal to 25, are presented. For comparison reasons, the rejection frequencies for the Christoffersen (1998)'s LR_{UC} and LR_{CC} test statistics are also reported. For a 5% coverage rate and whatever the choice of m, the empirical size of the J_{CC} test is close to the nominal size, even for small sample sizes. On the contrary, we verify that the LR_{CC} tests is over-sized. For a 1% VaR, the J_{CC} test is also well sized, whereas the LR_{CC} test seems to be undersized in small samples, although it size converges to the nominal one as T increases. The performance of our test is quite remarkable, since under the null in a sample with $T = 250$ and a coverage rate equal to 1%, the expected number of violations lies between two and three.

The same type of results can be observed when the block size N is increased. The rejection frequencies of the Monte-Carlo experiments for the $J_{CC}(m)$ GMM-based test statistic, both for a coverage rate of 5% and of 1% and for a block of size 100 are reported in table 2. In this case, our statistic is based on fewer observations, but still has a good size, close to the nominal size of 5%.

[Insert Table 2]

It is important to note that these rejection frequencies are only calculated for the simulations providing a LR test statistic. Indeed, for realistic sample size (for instance $T = 250$) and a coverage rate of 1%, many simulations do not deliver a LR statistic. The LR_{UC} tests can be computed without any restrictions. But, the LR_{CC} test statistic is computable only when the independence test (J_{IND}) is feasible, that is, if there is at least one violation in the sample. Thus, at a 1% coverage rate for which the scarcity of violations is more obvious, a large sample size is required in order to compute this test statistic. The fraction of samples for which a test is feasible is reported for each sample size, both for the size and power experiments (at 5% and 1% coverage rate), are reported in table 3. By contrast, our GMM-based test can always be computed as long as the number of moment conditions m is inferior or equal to the block size N. It is one of the advantage of our approach.

[Insert Table 3]

4.2 Empirical Power Analysis

We now investigate the empirical power of our GMM test, especially in the context of risk management. As previously mentioned, Value-at-Risk (VaR) forecasts can be interpreted as one-sided and open forecast intervals. More formally, let us consider an interval $CI_{t|t-1}(\alpha) = [-\infty, VaR_{t|t-1}(\alpha)]$, where $VaR_{t|t-1}(\alpha)$ denotes the conditional VaR obtained for a coverage (or risk) equal to $\alpha\%$. As usual in the backtesting literature, our power experiment consists is based on a particular DGP for financial returns and a method to compute VaR forecast. This method has to be chosen to produce invalid VaR forecasts according to Christofersen's hypotheses.

Following Berkowitz et al. (2010), we assume that returns r_t are issued from a simple t-GARCH model with an asymmetric leverage effect:

$$r_t = \sigma_t z_t \sqrt{\frac{\nu - 2}{\nu}}, \tag{19}$$

where z_t is an $i.i.d.$ series from Student's t-distribution with ν degrees of freedom, and where the conditional variance σ_t^2 is given:

$$\sigma_t^2 = \omega + \gamma \sigma_{t-1}^2 \left(\sqrt{\frac{\nu - 2}{\nu}} z_{t-1} - \theta \right)^2 + \beta \sigma_{t-1}^2. \tag{20}$$

Once the returns series has been generated [2], a method of VaR forecasting must be selected. Obviously, this choice has deep implications in terms of power performance for the interval forecast evaluation tests. Thus, the designated method should produce VaR forecasts that violate the efficiency hypothesis (the UC and/or IND assumptions). Here, we have chosen the same method as in Berkowitz et al. (2010), i.e. the historical simulation (HS), with a rolling window T_e equal to 250. Formally, we define the HS-VaR as following:

$$VaR_{t|t-1}(\alpha) = Percentile \left(\{r_i\}_{i=t-T_e}^{t-1}, 100\alpha \right). \tag{21}$$

For each simulation, a violation sequences $\{I\}_{t=1}^{T}$ is then constructed, by comparing the *ex-ante* $VaR_{t|t-1}(\alpha)$ forecasts to the *ex post* returns r_t. Next, the partial sums sequence $\{y_h\}_{h=1}^{H}$ is computed, for a given block size N, by summing the corresponding I_t observations (see section 3.1). Based on this sequence, the J_{CC} test statistics are then implemented for different number of

[2] The coefficients of the model are parametrized as in Berkowitz et al. (2010): $\gamma = 0.1$, $\theta = 0.5$, $\beta = 0.85$, $\omega = 3.9683e^{-6}$ and $d = 8$. At the same time, ω has been chosen so as to be consistent with a 0.2 annual standard deviation. Additionally, the global parametrization corresponds to a daily volatility persistence of 0.975.

moment conditions and sample sizes T ranging from 250 to 1500. For comparison, both LR_{UC} and LR_{CC} statistics are also computed for each simulation. The rejection frequencies, at a 5% nominal size, are based on 10,000 simulations. In order to control for size distortions between LR and J_{CC} tests, we use the Dufour (2006)'s Monte-Carlo method (see appendix B).

[Insert Tables 4 and 5]

Tables 4 and 5 report the corrected power of the J_{UC}, $J_{CC}(m)$, LR_{UC} and LR_{CC} tests for different sample sizes T, in the case of a 5% and a 1% coverage rate, both for a block size $N = 25$ and $N = 100$. We can observe that the two GMM-based tests (J_{UC} and J_{CC}) have good small sample power properties, whatever the sample size T and the block size N considered. Additionally, our test is proven to be quite robust to the choice of the number of moment conditions m. Nevertheless, in our simulations it appears that the optimal power of our GMM-based test is reached when considering two moment conditions ($m = 2$). More precisely, for a 5% coverage rate, a sample size $T = 250$ and a block size $N = 25$, the power of our $J_{CC}(2)$ test statistic is approximately two times the power of the corresponding LR test. In the case of an $\alpha = 1\%$ coverage rate, the power of our $J_{CC}(2)$ test remains by 30% higher than the one of the LR test. On the contrary, our unconditional coverage J_{UC} test does not outperform the LR test. This result is logical, since both exploit approximately the same information, *i.e.* the frequency of violations.

The choice of the block size N has two opposite effects on the empirical power. An increase in the block size N leads to an decrease of the length of the sequence $\{y_h\}_{h=1}^{H}$ used to compute the J-statistic, and then leads to a decrease of its empirical power. On the contrary, an increase in the block size N allows taking into account a larger number of indicator variables I_t in each block. Consequently, the test is more efficient in detecting deviations from UC and/or IND assumptions inside each block. Figure 2 displays the Dufour's corrected empirical power of the $J_{CC}(2)$ statistic as a function of the sample size T, for three values $(2, 25$ and $100)$ of the block size N. We note that, whatever the sample size, the power for a block size $N = 100$ is always lesser than that obtained for a block size equal to 25. In the same time, the power with $N = 100$ is always larger than that with $N = 2$. In order to get a more precise idea of the link between the power and the block size N, the Figure 3 displays the Dufour's corrected empirical power of the $J_{CC}(2)$ statistic as a function of the block size N, for three values $(250, 750$ and $1500)$ of the sample size T. The highest corrected power corresponds to block sizes between 20 and 40, that is why we recommend a value of $N = 25$ for applications. Other simulations based on Bernoulli trials with a false coverage rate(available upon request) confirm this choice.

[Insert Figures 3 and 4]

Thus, our new GMM-based interval forecasts evaluation tests seems to perform better both in terms of size and power than the traditional LR ones.

5 An Empirical Application

To illustrate the finite sample properties of our test, we propose to an empirical application based on two series of daily returns, namely the SP500 (from 05 January 1953 to 19 December 1957) and the Nikkei (from 27 January 1987 to 21 February 1992). The baseline idea is to select some periods and assets for which the linearity assumption is strongly rejected by standard specification tests. Then, we use (at wrong) a linear model to produce a sequence of invalid interval forecasts. The issue is then to check if our evaluation tests are able to reject the nulls of UC, IND and/or CC.

Here we use the nonlinearity test recently proposed by Harvey and Laybourne (2007). This takes into account both an ESTAR or LSTAR alternative hypothesis, and has very good small sample properties. For the considered periods, the conclusion of the test (see Table 5) are clear: the linearity assumption is strongly rejected for both assets. For the SP500 (resp. Nikkei), the statistic is equal to 24.509 (resp. 89.496) with a p-value less than 0.001. Newt, we use simple autoregressive linear models AR(1) to produce forecasts and interval forecasts at an horizon $h = 1, 5$ or 10 days. More precisely, each model is estimated on the first 1,000 in sample observations, while continuous and symmetrical confidence intervals are computed for each sequence of 250 out-of-sample observations both at a 5% and 1% coverage rate.

[Insert Tables 6 and 7]

Tables 6 and 7 report the main results of the interval forecast tests, based on a block size N equal to 25. It appears that for the SP500 index (see Table 6) our GMM-based test always rejects the CC hypothesis and thus, the validity of the forecasts. At the same time, the LR_{CC} test does not reject this hypothesis for a 5% coverage rate. Moreover, when considering a 1% coverage rate, both CC tests succeed in rejecting the null hypothesis. Still, further clarifications are required. Both the UC and IND hypothesis are rejected when using GMM-based tests, whereas the only assumption rejected by the LR tests is the UC one. Similar results are obtained for the Nikkei series (see Table 7). Thus, the two series of interval forecasts are characterized by clusters of violations detected only by our GMM-based test. On the contrary, the LR_{IND} test appears not to be powerful enough to reject the independence assumption. This analysis proves that our evaluation tests for interval forecasts has interesting properties for applied econometricians, especially when they have to evaluate the validity of interval forecasts on short samples.

6 Conclusion

This paper proposes a new evaluation framework of interval forecasts based on simple J-statistics. Our test is model free and can be applied to intervals and/or HDR forecasts, potentially discontinuous and/or asymmetric. The underlying idea is that if the interval forecast is correctly specified, then the sum of the violations should be distributed according to Binomial distribution with a success probability equal to the coverage rate. So, we adapt the GMM framework proposed by Bontemps (2006) in order to test for this distributional assumption that corresponds to the null of interval forecast validity.

More precisely, we propose an original approach, derived from the subsampling methodology that transforms the violation series into a series of partial sums of violations defined on sub-block. Under the null of validity, these partial sums are distributed according a Binomial distribution.

Our approach has several advantages. First, all three hypotheses of unconditional coverage, independence and conditional coverage can be tested independently. Second, these testes are easy to implement and robust to distributional parameter's uncertainty. Third, Monte-Carlo simulations show that all our GMM-based tests outperform the LR ones in terms of power, especially in small samples and for a 5% coverage rate (95% interval forecasts), which are the most interesting cases from a practical viewpoint. Assessing the impact of the estimation risk for the parameters of the model that generated the HDR or the interval forecasts (and not for the distributional parameters) on the distribution of the GMM test-statistic by using a subsampling approach or a parametric correction is left for future research.

Bibliography

References

[1] Berkowitz, J., Christoffersen, P., Pelletier, D., (2010). Evaluating Value-at-Risk Models with Desk-Level Data, forthcoming in *Management Science*.

[2] Bontemps, C., (2006). Moment-based tests for discrete distributions, *Working Paper TSE*.

[3] Bontemps, C., and Meddahi, N., (2005). Testing normality: a GMM approach, *Journal of Econometrics,* 124, 149-186.

[4] Candelon, B., Colletaz, G., Hurlin, C., and Tokpavi, S., (2010). Backtesting Value-at-Risk: a GMM duration-based test, forthcoming in *Journal of Financial*

Econometrics.

[5] Chatfield, C., (1993). Calculating interval forecasts, *Journal of Business and Economic Statistics*, 11, issue 2, 121-135.

[6] Christoffersen, F.P., (1998). Evaluating interval forecasts, *International Economic Review*, 39, 841-862.

[7] Clements, M.P., Taylor, N. (2002). Evaluating interval forecasts of high-frequency financial data, *Journal of Applied Econometrics*, 18, Issue 4, 445 - 456.

[8] Colletaz G. and Hurlin C. (2005), Modèles non linéaires et prévision, *Rapport Institut CDC pour la recherche*, 106 pages.

[9] Dufour, J.-M., (2006). Monte Carlo tests with nuisance parameters: a general approach to finite sample inference and nonstandard asymptotics, *Journal of Econometrics*, 127, issue 2, 443-477.

[10] Hansen, L.P., (1982). Large sample properties of Generalized Method of Moments estimators, *Econometrica*, 50, 1029-1054.

[11] Harvey, D. I., and Leybourne, S. J., (2007), Testing for time series linearity, *Econometrics Journal,* 10, 149-165.

[12] Hyndman, R.J., (1995). Highest-density forecast regions for non-linear and non-normal time-series models, *Journal of Forecasting,* 14, 431-441.

[13] Politis, D.N., Romano, J.P., and Wolf, M., (1999), Subsampling, Springer-Verlag, New-York.

[14] Teräsvirta, T., (2006), Forecasting economic variables with non linear models, in Handbook of Economic Forecasting, G. Elliott, C.W.J. Granger and A. Timmermann editors, Elsevier, volume 1, Chapter 8, 413-457.

[15] Wallis, K.F., (2003). Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts, *International Journal of Forecasting*, 19, 165-175.

Appendix A: Proof of the robustness of J_{IND} to parameter uncertainty

According to Bontemps (2006), if an orthonormal polynomial is orthogonal to the score, this polynomial is robust to parameter uncertainty. In our case, we test the independence distribution of a series y_h drawn from a *i.i.d.* $B(N,p)$ distribution. Therefore, the probability distribution function of y_h has the following form:

$$f_D(N,\beta) = \frac{N!}{y_h!(N-y_h)!}\beta^{y_h}(1-\beta)^{(N-y_h)}. \tag{B.1}$$

Consequently, the score function can be calculated as:

$$\frac{\partial ln\ f_D(N,\beta)}{\partial \beta} = \frac{1}{f_D(N,\beta)}\frac{\partial f_D(N,\beta)}{\partial \beta} \tag{B.2}$$

$$= -\frac{N\beta - y_h}{\beta(1-\beta)}. \tag{B.3}$$

Let us recall that first moment condition takes the form of:

$$P_1^{(N,\alpha)}(y_h) = \frac{\alpha N - y_h}{\sqrt{\alpha(1-\alpha)N}}. \tag{B.4}$$

Based on equations B.3 and B.4, it is obvious that the score is proportional to first Krawtchouk polynomial:

$$\frac{\partial\ ln\ f_D(N,\beta)}{\partial \beta} = -P_1^{(N,\alpha)}(y_h)\sqrt{\frac{N}{\beta(1-\beta)}}. \tag{B.5}$$

Since the other m polynomials are obtained using the recursive formula developed in equation B.5, it is obvious that they are also proportional to the score. More precisely, if we multiply equation B.5 by any polynomial of order greater than 2, and by applying the mathematical expectation operator, we obtain:

$$\mathbb{E}\left[\frac{\partial ln\ f_D(N,\beta)}{\partial \beta}P_i^{(N,\alpha)}(y_h)\right] = -\sqrt{\frac{N}{\beta(1-\beta)}}\mathbb{E}\left[P_1^{(N,\alpha)}(y_h)P_i^{(N,\alpha)}(y_h)\right] = 0,\ \forall i \geq 2. \tag{B.6}$$

Thus, these robust moments are actually identical to the initial moments, which means that we can create a GMM J_{IND} statistic based on estimated β with no modification in the asymptotic distribution, as long as the estimated $\hat{\beta}$ is the square-T-root-consistent estimator of the true β.

Appendix B: Dufour (2006) Monte-Carlo Corrected Method

To implement MC tests, first generate M independent realizations of the test statistic, say S_i, $i = 1, \ldots, M$, under the null hypothesis. Denote by S_0 the value of the test statistic obtained for the original sample. As shown by Dufour (2006) in a general case, the MC critical region is obtained as $\hat{p}_M(S_0) \leq \eta$ with $1 - \eta$ the confidence level and $\hat{p}_M(S_0)$ defined as

$$\hat{p}_M(S_0) = \frac{M \; \hat{G}_M(S_0) + 1}{M + 1}, \tag{22}$$

where

$$\widehat{G}_M(S_0) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}(S_i \geq S_0), \tag{23}$$

when the ties have zero probability, *i.e.* $\Pr(S_i = S_j) \neq 0$, and otherwise,

$$\widehat{G}_M(S_0) = 1 - \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}(S_i \leq S_0) + \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}(S_i = S_0) \times \mathbb{I}(U_i \geq U_0). \tag{24}$$

Variables U_0 and U_i are uniform draws from the interval $[0, 1]$ and $\mathbb{I}(.)$ is the indicator function. As an example, for MC tests procedure applied to the test statistic $S_0 = J_{CC}(m)$, we just need to simulate under H_0, M independent realizations of the test statistic (*i.e.*, using durations constructed from independent Bernoulli hit sequences with parameter α) and then apply formulas (22 to 24) to make inference at the confidence level $1 - \eta$. Throughout the paper, we set M at $9,999$.

Table 1. Empirical size (block size $N = 25$)

Coverage rate 5%							
T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	10	0.0389	0.0537	0.0447	0.0390	0.0560	0.0857
500	20	0.0483	0.0505	0.0517	0.0440	0.0545	0.0906
750	30	0.0432	0.0508	0.0589	0.0557	0.0532	0.1017
1000	40	0.0512	0.0556	0.0565	0.0469	0.0534	0.1103
1250	50	0.0504	0.0463	0.0528	0.0452	0.0441	0.1302
1500	60	0.0511	0.0474	0.0516	0.0432	0.0518	0.1115
Coverage rate 1%							
T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	10	0.0450	0.0545	0.0545	0.0461	0.0150	0.0216
500	20	0.0312	0.0654	0.0632	0.0535	0.0663	0.0195
750	30	0.0670	0.0626	0.0698	0.0681	0.0431	0.0326
1000	40	0.0394	0.0527	0.0547	0.0811	0.0578	0.0376
1250	50	0.0437	0.0590	0.0446	0.0429	0.0621	0.0360
1500	60	0.0476	0.0486	0.0434	0.0274	0.0551	0.0428

Note: Under the null hypothesis, the violations are i.i.d. and follows a Bernoulli distribution. The results are based on 10,000 replications. For each sample, we provide the percentage of rejection at a 5% level. Jcc(m) denotes the GMM based conditional coverage test with m moment conditions. Juc denotes the unconditional coverage test obtained for m=1. LRcc (resp. LRuc) denotes the Christoffersen's conditional (resp. unconditional) coverage test. T denotes the sample size of the sequence of interval forecats violations I_t, while H=[T/N] denotes the number of block (size N=25) used to define the partial sums (y_h) of violations.

Table 2. Empirical size (block size $N = 100$)

T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
Coverage rate 5%							
250	2	0.0313	0.0631	0.0511	0.0436	0.0587	0.0851
500	5	0.0515	0.0556	0.0635	0.0687	0.0556	0.0967
750	7	0.0426	0.0526	0.0601	0.0743	0.0514	0.0956
1000	10	0.0497	0.0540	0.0612	0.0672	0.0530	0.1092
1250	12	0.0516	0.0522	0.0592	0.0637	0.0422	0.1284
1500	15	0.0483	0.0508	0.0647	0.0636	0.0484	0.1088
Coverage rate 1%							
T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	2	0.0579	0.0451	0.0451	0.0451	0.0138	0.0222
500	5	0.0326	0.0414	0.0369	0.0387	0.0636	0.0215
750	7	0.0346	0.0510	0.0447	0.0402	0.0432	0.0304
1000	10	0.0342	0.0505	0.0590	0.0511	0.0543	0.0348
1250	12	0.0548	0.0540	0.0597	0.0592	0.0604	0.0340
1500	15	0.0478	0.0539	0.0513	0.0452	0.0536	0.0396

Note: Under the null hypothesis, the violations are i.i.d. and follows a Bernoulli distribution. The results are based on 10,000 replications. For each sample, we provide the percentage of rejection at a 5% level. Jcc(m) denotes the GMM based conditional coverage test with m moment conditions. Juc denotes the unconditional coverage test obtained for m=1. LRcc (resp. LRuc) denotes the Christoffersen's conditional (resp. unconditional) coverage test. T denotes the sample size of the sequence of interval forecats violations I_t, while H=[T/N] denotes the number of block (size N=100) used to define the partial sums (y_h) of violations.

Table 3. Feasibility ratios

		Size simulations ($N = 25$)					
		5% coverage rate			1% coverage rate		
T	H	J_{CC}	LR_{UC}	LR_{CC}	J_{CC}	LR_{UC}	LR_{CC}
250	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9185
500	20	1.0000	1.0000	1.0000	1.0000	1.0000	0.9940
750	30	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995
1000	40	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		Power simulations ($N = 25$)					
		5% coverage rate			1% coverage rate		
T	H	J_{CC}	LR_{UC}	LR_{CC}	J_{CC}	LR_{UC}	LR_{CC}
250	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9042
500	20	1.0000	1.0000	1.0000	1.0000	1.0000	0.9978
750	30	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
1000	40	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Note: the fraction of samples for which a test is feasible is reported for each sample size, both for the size and power tests at a 5% and respectively 1% coverage rate. J_{CC} represents the conditional coverage GMM-based test, while LR_{UC} and LR_{CC} are Christoffersen (1998)'s unconditional and conditional coverage LR tests. Note that for J_{CC} the feasibility ratios are independent of the number of moment conditions m. All results are based on 10,000 simulations.

Table 4. Empirical Power (block size $N = 25$)

				Coverage rate 5%			
T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	10	0.2446	0.5140	0.5051	0.4645	0.2164	0.2521
500	20	0.1765	0.7045	0.6967	0.6793	0.1494	0.2739
750	30	0.1502	0.8314	0.8271	0.8113	0.1159	0.3267
1000	40	0.1544	0.9079	0.8997	0.8916	0.1216	0.3886
1250	50	0.1455	0.9533	0.9452	0.9368	0.1251	0.4445
1500	60	0.1664	0.9724	0.9664	0.9603	0.1289	0.4997
				Coverage rate 1%			
T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	10	0.2449	0.3693	0.3854	0.3860	0.1783	0.2659
500	20	0.2430	0.5039	0.5394	0.5385	0.1501	0.2454
750	30	0.2617	0.6448	0.6576	0.6340	0.2101	0.2824
1000	40	0.3392	0.7220	0.7358	0.7125	0.2090	0.3868
1250	50	0.3717	0.7932	0.7926	0.7943	0.2711	0.4384
1500	60	0.4070	0.8455	0.8472	0.8331	0.3340	0.4942

Note: Power simulation results are provided for different sample sizes T and number of blocks H, both at a 5% and 1% coverage rate. $J_{CC}(m)$ denotes the conditional coverage test with m moment conditions, J_{UC} represents the unconditional coverage test for the particular case when $m = 1$, and LR_{UC} and LR_{CC} are the unconditional and respectively conditional coverage tests of Christoffersen (1998). The results are obtained after 10,000 simulations by using Dufour (2005)'s Monte-Carlo procedure with ns=9999. The rejection frequencies are based on a 5% nominal size.

Table 5. Empirical Power (block size $N = 100$)

Coverage rate 5%							
T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	2	0.2892	0.4096	0.4373	0.4356	0.2379	0.2718
500	5	0.1498	0.6206	0.6372	0.6116	0.1546	0.2783
750	7	0.1261	0.7082	0.7163	0.6999	0.1163	0.3357
1000	10	0.1192	0.8013	0.8086	0.8021	0.1160	0.3779
1250	12	0.1291	0.8647	0.8677	0.8630	0.1213	0.4343
1500	15	0.1364	0.9108	0.9097	0.9047	0.1304	0.5043
Coverage rate 1%							
T	H	J_{UC}	$J_{CC}(2)$	$J_{CC}(3)$	$J_{CC}(5)$	LR_{UC}	LR_{CC}
250	2	0.1902	0.2800	0.2800	0.2908	0.1892	0.2750
500	5	0.2233	0.4493	0.4587	0.4461	0.1507	0.2487
750	7	0.2572	0.5401	0.5565	0.5448	0.2064	0.2871
1000	10	0.2855	0.6392	0.6480	0.6511	0.2058	0.3802
1250	12	0.3422	0.7146	0.7141	0.7174	0.2711	0.4339
1500	15	0.3774	0.7760	0.7712	0.7708	0.3341	0.4984

Note: Power simulation results are provided for different sample sizes T and number of blocks H, both at a 5% and 1% coverage rate. $J_{CC}(m)$ denotes the conditional coverage test with m moment conditions, J_{UC} represents the unconditional coverage test for the particular case when $m = 1$, and LR_{UC} and LR_{CC} are the unconditional and respectively conditional coverage tests of Christoffersen (1998). The results are obtained after 10,000 simulations by using Dufour (2005)'s Monte-Carlo procedure with ns=9999. The rejection frequencies are based on a 5% nominal size.

Table 6. Interval Forecast Evaluation (SP500)

	Coverage rate 5%					
	GMM-based tests			LR tests		
Horizon	J_{UC}	$J_{IND}(2)$	$J_{CC}(2)$	LR_{UC}	LR_{IND}	LR_{CC}
1	2.5263 (0.1120)	11.612 (0.0006)	29.493 (<0.0001)	2.4217 (0.1197)	3.3138 (0.0687)	5.8816 (0.0528)
5	4.4912 (0.0341)	10.615 (0.0011)	37.604 (<0.0001)	4.0607 (0.0439)	7.5661 (0.0059)	11.787 (0.0028)
10	2.5263 (0.1120)	19.605 (<0.0001)	46.040 (<0.0001)	2.4217 (0.1197)	3.3138 (0.0687)	5.8816 (0.0528)
	Coverage rate 1%					
	GMM-based tests			LR tests		
Horizon	J_{UC}	$J_{IND}(2)$	$J_{CC}(2)$	LR_{UC}	LR_{IND}	LR_{CC}
1	109.09 (<0.0001)	11.612 (0.0006)	2072.4 (<0.0001)	49.234 (<0.0001)	3.3138 (0.0687)	52.693 (<0.0001)
5	134.68 (<0.0001)	10.615 (0.0011)	2658.3 (<0.0001)	57.475 (<0.0001)	7.5661 (0.0059)	65.201 (<0.0001)
10	109.09 (<0.0001)	19.605 (<0.0001)	2714.6 (<0.0001)	49.234 (<0.0001)	3.3138 (0.0687)	52.693 (<0.0001)

Note: 250 out of sample forecasts of the SP500 index (from 20/12/1956 to 19/12/1957) are computed for three different horizons (2, 5 and 10) both at a 5% and 1% coverage rate. The evaluation results of the corresponding interval forecasts are reported both for our GMM-based tests and Christoffersen (1998)'s LR tests. For this objective, a block size N=25 was used. For all tests, the numbers in the parentheses denote the corresponding p-values.

Table 6. Interval Forecast Evaluation (SP500)

	Coverage rate 5%					
	GMM-based tests			LR tests		
Horizon	J_{UC}	$J_{IND}\,(2)$	$J_{CC}\,(2)$	LR_{UC}	LR_{IND}	LR_{CC}
1	2.5263 (0.1120)	3.9132 (0.0479)	12.060 (0.0024)	1.7470 (0.1863)	0.2521 (0.6156)	2.1382 (0.3433)
5	1.7544 (0.1853)	3.8728 (0.0491)	9.6337 (0.0081)	1.1744 (0.2785)	0.4005 (0.5268)	1.7072 (0.4259)
10	1.7544 (0.1853)	3.8728 (0.0491)	9.6337 (0.0081)	1.1744 (0.2785)	0.4005 (0.5268)	1.7072 (0.4259)
	Coverage rate 1%					
	GMM-based tests			LR tests		
Horizon	J_{UC}	$J_{IND}\,(2)$	$J_{CC}\,(2)$	LR_{UC}	LR_{IND}	LR_{CC}
1	109.09 (0.0000)	3.9132 (0.0479)	1279.3 (0.0000)	45.258 (<0.0001)	0.2521 (0.6100)	45.649 (<0.0001)
5	97.306 (0.0000)	3.8728 (0.0491)	1073.6 (0.0000)	41.384 (<0.0001)	0.4005 (0.5268)	41.916 (<0.0001)
10	97.306 (0.0000)	3.8728 (0.0491)	1073.6 (0.0000)	41.384 (<0.0001)	0.4005 (0.5268)	41.916 (<0.0001)

Note: 250 out of sample forecasts of the Nikkei index (from 27 January 1987 to 21 February 1992) are computed for three different horizons (2, 5 and 10) both at a 5% and 1% coverage rate. The evaluation results of the corresponding interval forecasts are reported both for our GMM-based tests and Christoffersen (1998)'s LR tests. For this objective, a block size N=25 was used. For all tests, the numbers in the parentheses denote the corresponding p-values.

Figure 2. Corrected power of the $J_{CC}(2)$ test statistic as function of the sample size T (coverage rate $\alpha = 5\%$)

Figure 3. Corrected power of the $J_{CC}(2)$ test statistic as function of the block size N (coverage rate $\alpha = 5\%$)